Benford-Newcomb Subsequences for Fraud Detection

Aaron Carl Smith

January 28, 2013

Abstract

Benford's law is frequently used to evaluate the likihood that data is misrepresentative. Typically statistical tests measure the likihood. Another method of employing Benford's law is to compare the frequency of leading digits to the probabilities of leading digits over a subset of the natural numbers. This paper proposes using the probabilities of leading digits from uniform, natural numbers to establish interval criteria for when to look more closely into the possibility of misrepresentative data.

Contents

1 Introduction

Benford's law gives a probability distribution for the frequency of the leading-digit of natural numbers. Simon Newcomb described the rule for decimal representation of natural numbers in 1881 [3], and Frank Benford generalized Newcomb's observations to any base in 1938 [1]. In 1995, Theodore Hill used the mantissa σ-algebra to further extend the leading-digit law to real numbers. The mantissa σ-algebra consists of sets of numbers with the same coefficient in scientific notation after truncation [2].

Definition 1.1 (Benford's Law)**.** *In base b, the probability that the leading digit of a real number is k is given by*

$$P(k) = log_b(1 + \tfrac{1}{k}),\ k \in \{1, 2, 3, \ldots, b-1\}. \tag{1.1}$$

In decimal representation (base 10), the probabilities of each the leading digits are given by

$$P(k) = log_{10}(1 + \tfrac{1}{k}),\ k \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}, \tag{1.2}$$

which approximately gives:

k	1	2	3	4	5	6	7	8	9
$P(k)$	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

The law goes further to say that the probability distribution of digits after the leading digit converges to uniform as the digit's position moves to the right [1, 2]. Benford's law does not apply to several types of numeric data, such as identification numbers.

2 Benford-Newcomb Subsequences

Consider the map f_b that sends natural numbers to their leading digits,

$$f_b : \mathbb{N} \to \{1, 2, 3, \ldots, b-1\}, x \mapsto floor(\tfrac{x}{b^{floor(log_b x)}}). \tag{2.1}$$

Let μ_N be the uniform probability measure on $\mathbb{N}$ where $\mu_N(k) = \frac{1}{N}\ \forall\ k \in \{1, 2, 3, \ldots, N\}$. Let's use μ_N to construct a probability measure of leading digits,

$$P_{bN}(k) = \mu_N(\{x \in \mathbb{N} | f_b(x) = k\}). \tag{2.2}$$

For a fixed base b and fixed leading digit k, consider the sequences $(P_{bN}(k))_{N=1}^{\infty}$; in general these sequences do not converge. The purpose of this paper is to propose using intervals of the form

$$[\liminf_{N\to\infty} P_{bN}(k), \limsup_{N\to\infty} P_{bN}(k)] \tag{2.3}$$

to identify possibly fraudulent data. If a data set's frequency of leading digits, in base b representation, is not contained in these intervals, then look further into the possibility of tamper data. For $N > b$, with respect to N the local minimums are of the form

$$P_{bN}(k) = \tfrac{1+b+b^2+\ldots+b^{\alpha-1}}{kb^{\alpha}-1},\ N = kb^{\alpha} - 1 \tag{2.4}$$

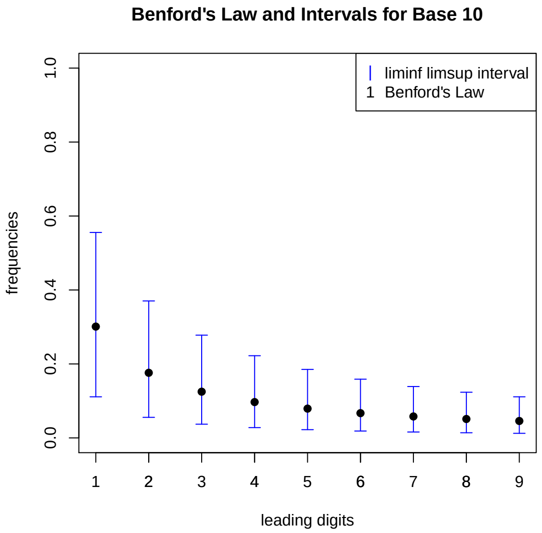
and the local maximums are of the form

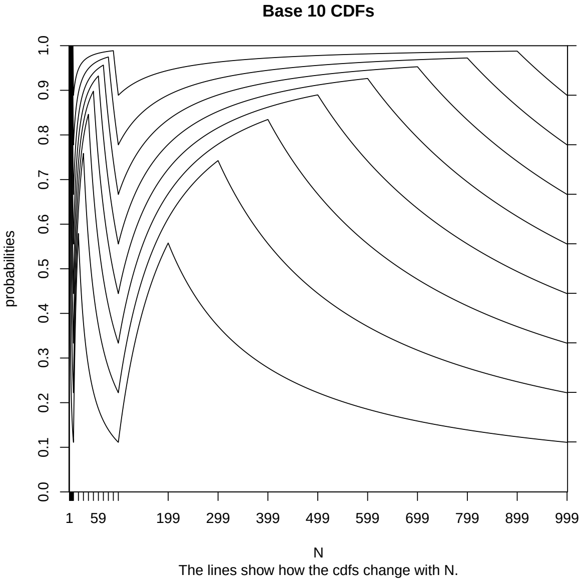
$$P_{bN}(k) = \tfrac{1+b+b^2+\ldots+b^{\alpha}}{(k+1)b^{\alpha}-1},\ N = (k+1)b^{\alpha} - 1. \tag{2.5}$$

Thus if the frequencies of a data set's leading digits are not within

$$[\tfrac{1}{k(b-1)}, \tfrac{b}{(k+1)(b-1)}], \tag{2.6}$$

further inquiry is called for. The advantage of the interval method is that one may use it to quickly screen data.





The figures were constructed with R [4].

References

[1] F. Benford, *The law of anomalous numbers*, Proceedings of the American Philosophical Society (1938), 551–572.

[2] Theodore P. Hill, *A statistical derivation of the significant-digit law*, Statist. Sci. **10** (1995), no. 4, 354–363. MR 1421567 (98a:60021)

[3] Simon Newcomb, *Note on the Frequency of Use of the Different Digits in Natural Numbers*, Amer. J. Math. **4** (1881), no. 1-4, 39–40. MR 1505286

[4] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012, ISBN 3-900051-07-0.